Method for training Neural Networks

BACKGROUND OF THE INVENTION

[0001] The invention relates to a method for training neural networks, to a method for prognosis by means of neural networks and to a system for determining a prognosis value and its error.

[0002] A large number of possible applications for neural networks are known from the prior art. Neural networks are used for data-driven modelling, for example for physical, biological, chemical and technical processes and systems, cf. Babel W.: Possible uses of neural networks in industry: pattern recognition with the aid of supervised learning methods - with examples from traffic and medical technology, Expert Verlag, Renningen-Malmsheim, 1997. In particular, the fields in which neural networks can be used include process optimization, image processing, pattern recognition, robot control and medical technology.

[0003] Before a neural network can be used for the purposes of prognosis or optimization, it has to be trained. The weightings of the neurons are in this case usually adjusted by an iterative method with the aid of training data, cf. Bärmann F.: Process modelling: modelling of continuous systems with neural networks, the NN-tool Internet site, www.baermann.de and Bärmann F.: Neural networks. Lecture text. Technical College of Gelsenkirchen, School of Physical Technology, Department of Neuro-Information Technology, 1998.

[0004] DE 195 31 967 discloses a method for training a neural network with the nondeterministic behaviour of a technical system. The neural network is in this case incorporated into a control loop so that, as it is output quantity, the neural network outputs a manipulated variable to the technical system and the technical system generates, from the manipulated variable delivered by the neural network, controlled variable which is delivered to the neural network as an input quantity. Noise with a

known noise distribution is superimposed on the controlled variable before it is delivered to the technical system.

[0005] Other methods for training neural networks are known from DE 692 28 412 T2 and DE 198 38 654 C1.

[0006] A method for estimating the prognosis error is known from EP 0 762 245 B1. Here, a plurality of neural networks having different training parameters (e.g. different initialisation) are trained with the original data. The prognosis error is obtained by comparing the discrepancies of the prognosed quantities. A disadvantage of this method is that the estimation of the prognosis error is not influenced by information about the measurement accuracy of the measurement data used for the training.

[0007] A method for estimating the reliability of a prognosis output by a neural network is furthermore known from the prior art: Protzel P., Kindermann L,. Tagscherer M., Lewandowski A.: "Estimation of the reliability of neural network prognoses in process optimization", VDI Report No. 1526, 2000.

SUMMARY OF THE INVENTION

[0008] It is an aspect of the invention to provide an improved method for training neural networks, a method for prognosis by means of neural networks and a system for determining prognosis values and their errors.

[0009] The aspect of the invention is respectively achieved by the features of the independent patent claims. Preferred embodiments of the invention are specified in the dependent claims.

[0010] The invention makes it possible to take account of the fact that the training data for the training of a neural network have only a finite accuracy. If the training data are determined by a measurement technique, for example, then a measurement

accuracy can be specified for each datum. In order to take this limited measurement accuracy into account for training the neural network, further sets of training data records are generated. Each individual set of training data is obtained from the original data records by perturbing the data in the framework of the measurement accuracy. One neural network is then trained for each training data record generated in this way.

[0011] If input data are entered into the neural networks trained in this way, then each of these networks generates a prognosis. Due to the different training data, which were obtained by perturbation and were used for training the neural networks, these prognoses can differ from one another. A mean, which has a higher accuracy and reliability then the individual prognoses, is preferably formed from the various prognoses.

[0012] According to a preferred embodiment of the invention, an equidistributed random variable with an expectation of zero is added to the training data. The random variable is in this case selected so that the result of the addition of the output datum and the random variable lies within the range given by the measurement accuracy.

[0013] According to another preferred embodiment of the invention, a normally distributed random variable with an expectation of zero is added. The variance of the random variable is selected so that the result of the addition of the output datum and the random variable lies with a predetermined probability within the range given by the measurement accuracy. This predetermined probability is, for example, 95% or more.

[0014] According to another preferred embodiment of the invention, the reliability of the prognosis value is determined on the basis of how the prognoses determined by the neural networks differ. If the individual prognoses of the neural networks differ greatly from one another, then it can be inferred from this that the neural networks do not constitute a reliable model in the range of the current input data. In order to

assess the reliability, the standard deviation of the individual prognoses may for example be calculated. If the standard deviation exceeds a permissible measure, it is concluded from this that a reliable prognosis is not possible.

[0015] A particular advantage is that the present invention permits estimation of the prognosis error and therefore objective appraisal of the quality of the prognosis. According to a preferred embodiment, the standard deviation is used for estimating the prognosis error. For example, the standard deviation itself is used as a measure of the prognosis error. As an alternative, the prognosis error is calculated from the standard deviation with the aid of a monotonic function, for example by multiplying the standard deviation with a problem-dependent constant factor.

[0016] According to another preferred embodiment of the invention, a signal is output when the standard deviation of the prognoses lies above a predetermined threshold value. The signal may, for example, lead to a visual and/or acoustic output for a user. This is particularly advantageous when the neural networks are being used for a control process.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] Preferred embodiments of the invention will be explained in more detail below with reference to the drawings, in which:

Figure 1 shows a block diagram of a neural network,

Figure 2A and 2B show a schematic representation of a training data record for the neural network in Figure 1,

Figure 3 shows a flow chart of an embodiment of the method according to the invention for training neural networks and for determining a prognosis value,

Figure 4 shows a block diagram of a system according to the invention for determining a prognosis value.

DETAILED DESCRIPTION OF THE PRESENTLY PREFERRED EMBODIMENTS

[0018] Figure 1 shows a neural network 1 with the inputs A, B and C and with the output D. In the general case, the neural network 1 may have an arbitrary number of inputs and outputs. The neural network may furthermore contain rigorous model components. In this case, the term hybrid neural network is used.

[0019] A training data record 2 (cf. figure 2A) is used for training the neural network 1. The training data record 2 involves a series of measurement samples, which respectively consist of measurement values for the measurement quantities A, B, C and D. These measurement quantities are assigned to the corresponding input of the neural network.

[0020] The measurement values are, for example, obtained from experimental series which are carried out in order to determine training data. The measurement accuracies of the individual measurement quantities are specified in a table 3. The measurement quantity A is determined with a measurement accuracy of +/- w, the measurement quantity B is determined with a measurement accuracy of +/- x, the measurement quantity C is determined with a measurement accuracy of +/- y and the measurement quantity D is determined with a measurement accuracy of +/- z.

[0021] Figure 3 illustrates a preferred embodiment of the method according to the invention. In step 30, a number n of training data records is generated from the training data record 2 (cf. Figure 2).

[0022] A further training data record i, with i lying between 1 and n, is obtained in the basis of the training data record 2 by perturbing the individual measurement values of the measurement quantities A, B, C and D for each measurement sample

while taking the respective measurement accuracy of the relevant measurement quantity into account. This process, i.e. the perturbation the training data record 2, is carried out n times so as to obtain the n training data records.

[0023] For example, the perturbation of the measurement values of the training data record 2 is carried out by adding an equidistributed random number. A perturbed measurement value x' is obtained from the originally measured measurement value x as:

$$x' = x + \text{random number},$$

where the random number lies in the interval

[lower limit; upper limit].

[0024] The lower limit and the upper limit may in this case depend on the original measurement value x and the measurement accuracy. If the measurement accuracy was specified as a relative error for the measurement value x, for example, then the following limits may be selected:

lower limit = - relative measurement accuracy * x
upper limit = relative measurement accuracy * x

[0025] In the normal case, the lower limit is a value less than 0, whereas the upper limit is a value greater than 0. In this case, the perturbed measurement value x' lies within the range of the measurement accuracy. If a measurement value for the measurement quantity A is to be perturbed, for example, then a random variable is added to the measurement value actually determined for the measurement quantity A.

[0026] This random variable is a random number from the interval [lower limit; upper limit], in which case the lower limit and the upper limit may depend on the

measurement value x (see above) or the lower limit and the upper limit are fixed quantities.

[0027] As an alternative, the random number may also be normally distributed. In this case, the variance of the random number is selected so that the perturbed measurement value x' lies with a predetermined probability in the tolerance range predetermined by the measurement accuracy. The predetermined probability may, for example, be selected as $\geq 95\%$.

[0028] In step 31, a number n+1 of neural networks of the same type as the neural network 1 in Figure 1 are trained with the training data record 2 and the n further training data records obtained in this way. This provides n neural networks with the same input and output parameters, which have been trained with perturbed training data records based on the same training data record determined by a measurement technique.

[0029] In order to determine a prognosis value, in step 32 an input data record of parameter values is entered into the inputs A, B and C of the neural networks trained in step 31. In step 33, the neural networks respectively output a prognosis value at their outputs; in the exemplary case in question, there are hence n+1 prognosis values at the outputs D.

[0030] The mean and the standard deviation of the prognosis values output in step 33 are calculated in step 34. Step 35 tests whether the standard deviation lies above a threshold value. If it does, a warning is outputted in step 36. The output of a warning means that the prognosis of the neural networks for the current training data record is not sufficiently reliable.

[0031] If the standard deviation lies below the threshold value, however, then no such warning is outputted in step 37, the mean of the prognosis values is outputted as the result of the prognosis. In addition or as an alternative, the standard deviation is output as a measure of the prognosis error.

[0032] Figure 4 shows a block diagram of an embodiment of the system according to the invention. Elements of Figure 4 which correspond to elements of Figure 1 are denoted by the same reference numbers.

[0033] The system involves a number n+1 of neural networks 1, which are constructed identically in principle, i.e. each have the inputs A, B and C and with the output D.

[0034] The inputs A, B and C of the neural networks 1 are connected to an input module 4 via which an input parameter record, for which a prognosis is to be carried out, is entered into the neural networks 1.

[0035] Due to this input of input data, each of the neural networks 1 outputs a prognosis for the measurement quantity D at its output. These prognoses are entered into the evaluation module 5. The evaluation module 5 calculates the mean of the various prognoses for the measurement quantity D, as well as the standard deviation of the prognoses.

[0036] The evaluation module 5 is connected to a display unit 6 and to a comparator 7. The comparator 7 is connected to a threshold-value memory 8.

[0037] The evaluation module 5 outputs the calculated mean of the prognoses to the display unit 6, so that the prognosis value is displayed there. The evaluation module 5 furthermore outputs the standard deviation of the prognoses to the comparator 7. The comparator 7 compares the standard deviation with the threshold value stored in the threshold-value memory 8. If the standard deviation lies above the threshold value, then the comparator 7 outputs a signal which is displayed as warning information on the display unit 6. Hence, if a prognosis value is displayed on the display unit 6 together with warning information, this means that the spread of the individual prognoses of the neural networks 1 is relatively large, so that the resulting prognosis value is not regarded as sufficiently reliable.

[0038] The standard deviation may furthermore be output for a user via the display 6 as a measure of the prognosis error.